



Universiteit
Leiden
Governance and Global Affairs

udp UNIVERSIDAD
DIEGO PORTALES


ODISSEI
Open Data Infrastructure for
Social Science and Economic Innovations

Measuring Complexity and Reproducibility

A Comprehensive Benchmark of LLMs for Multilingual Policy Agenda Topic Annotation

Bastián González-Bustamante

Leiden University

✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation at the ODISSEI Conference for Social Sciences in the Netherlands 2025
Utrecht, Netherlands, November 4, 2025

Introduction

Research Overview

- ❑ **Benchmark study.** Testing 79 contemporary LLMs on zero-shot classification of policy topics.
- ❑ **Multilingual approach.** Covering Danish, Dutch, English, French, Hungarian, Italian, Portuguese and Spanish.
- ❑ **Performance analysis.** Evaluating how reasoning capabilities, model architecture and deployment affect accuracy.



Artwork by Leonardo Phoenix model

The Challenge of Policy Annotation



Artwork by DALL-E 3 model

- ⚖️ **Complex classification.** Assigning one of 21 major policy topics to legislative texts.
- 🌐 **Language barriers.** Working across eight different languages with varying resources.
- 📝 **Semantic nuance.** Requiring fine distinctions between related policy areas.
- ⌚ **Reproducibility concerns.** Balancing performance with scientific reproducibility.

Empirical Expectations

Research Hypotheses

💡 Reasoning Advantage Hypothesis

LLMs explicitly optimised for chain-of-thought (CoT) reasoning will achieve better performance in zero-shot policy agenda classification.

🔒 Openness Penalty Hypothesis

Proprietary, closed-source LLMs will outperform open-source LLMs on zero-shot policy agenda classification.

♻️ Reproducibility Penalty Hypothesis

LLMs deployed locally using a deterministic sampling will achieve lower performance in zero-shot policy agenda classification.

Methods

Ground-Truth Data

Denmark

15 101 bills
1953–2016

NLD

4 684 bills
1981–2009

UK

6 169 Acts
1911–2015

France

3 069 laws
1979–2013

Hungary

8 220 bills
1990–2022

Italy

4 554 laws
1983–2013

Brazil

2 449 laws
2003–2014

Spain

2 256 laws-decrees
1980–2018

Note. We split the samples in a proportion of 70/15/15 (stratified by major topic) for training, validation, and testing for future fine-tuning jobs. The samples correspond to ground-truth data of the Comparative Agendas Project.

79 LLMs

Run **620 times** under different conditions
(e.g., parameters, API/local, temperature,
datasets/language) for

- (1) **overall performance metrics**
- (2) **meta-analysis**

* GPT-5 and OSS (August 2025).



SOTA closed-source LLMs

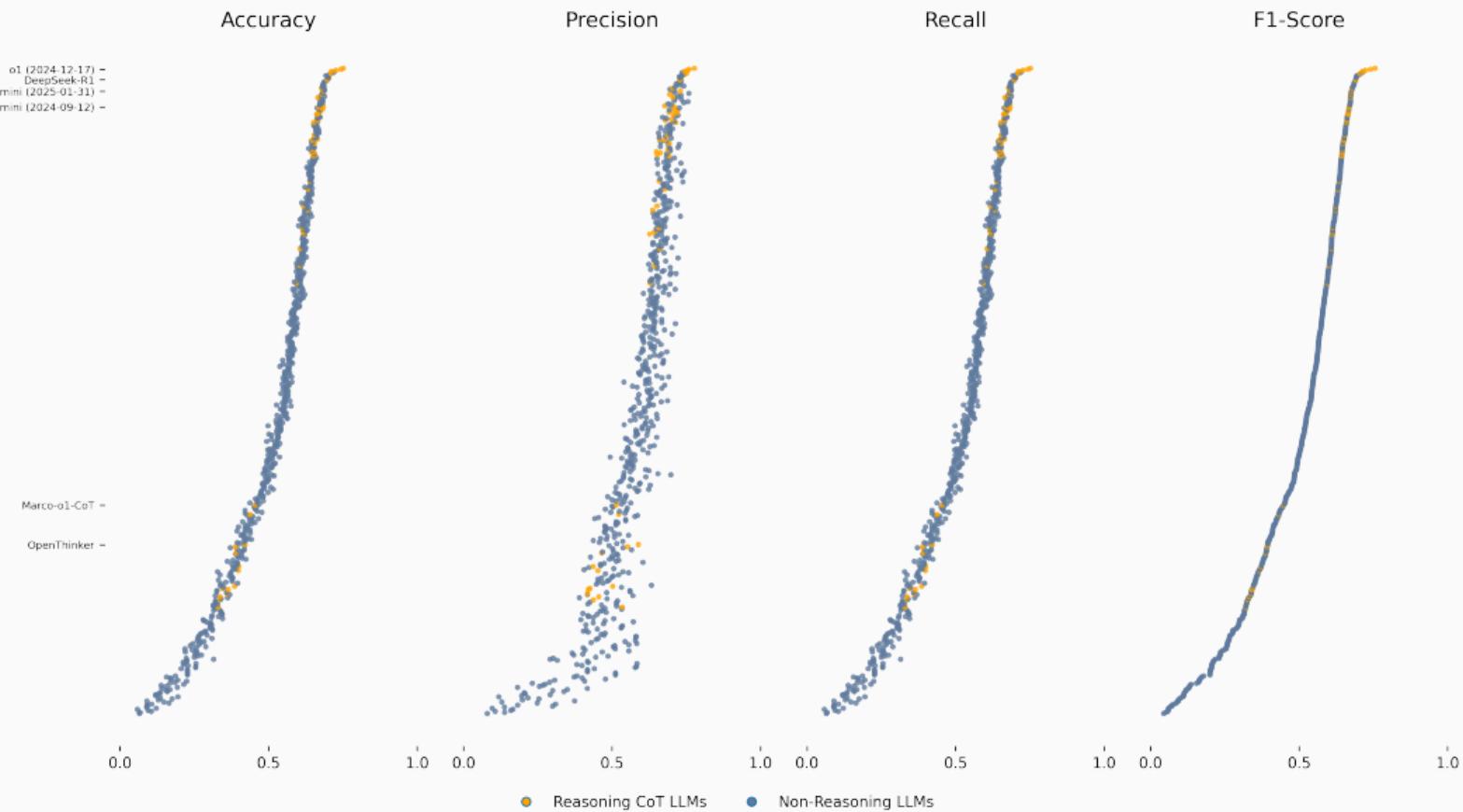
o4-mini, o3-mini, o1, GPT-4.1,
GPT-4.5-preview, Grok 3 Beta, Claude 3.7
Sonnet, among others



SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B),
Mistral 3.1 (24B), Llama 3.3 (70B),
DeepSeek-R1 (671B), DeepSeek-V3 (671B),
among others

Results



Top Performing Models

0.76

**Maximum
F1-Score**

Achieved by OpenAI
reasoning models

0.69

**Top Open-
Source Models**

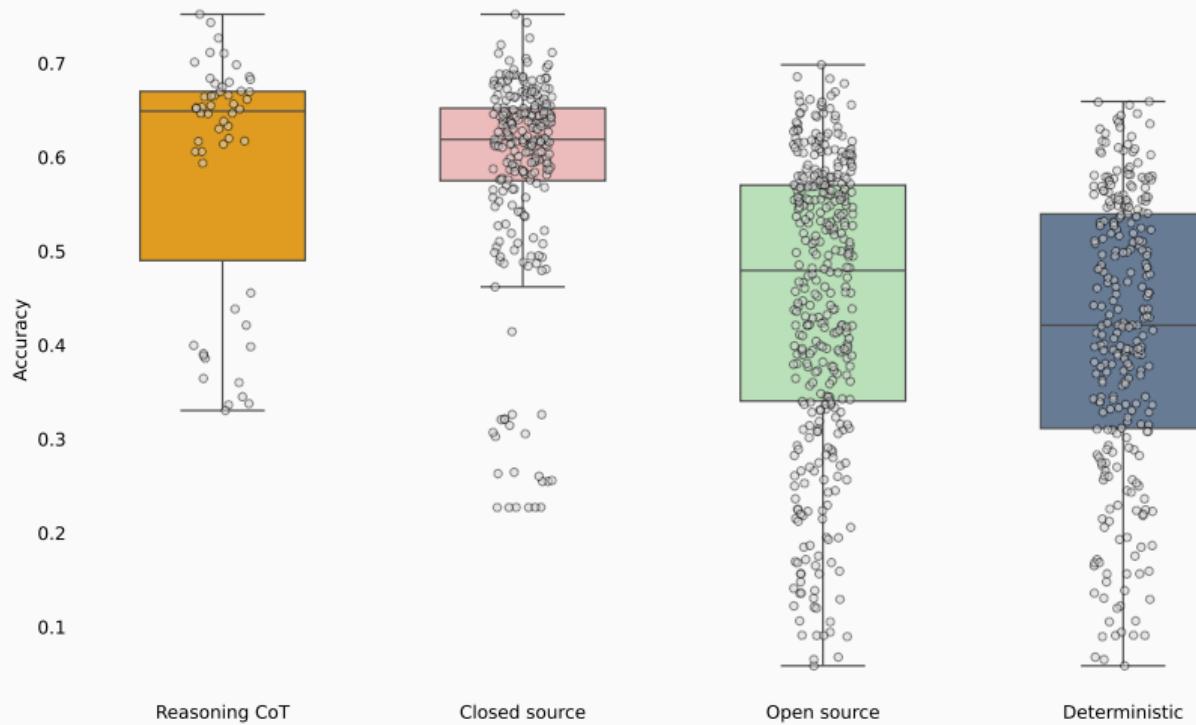
DeepSeek-R1 and V3 (671B)
and Llama 3.1 (405B)

0.54

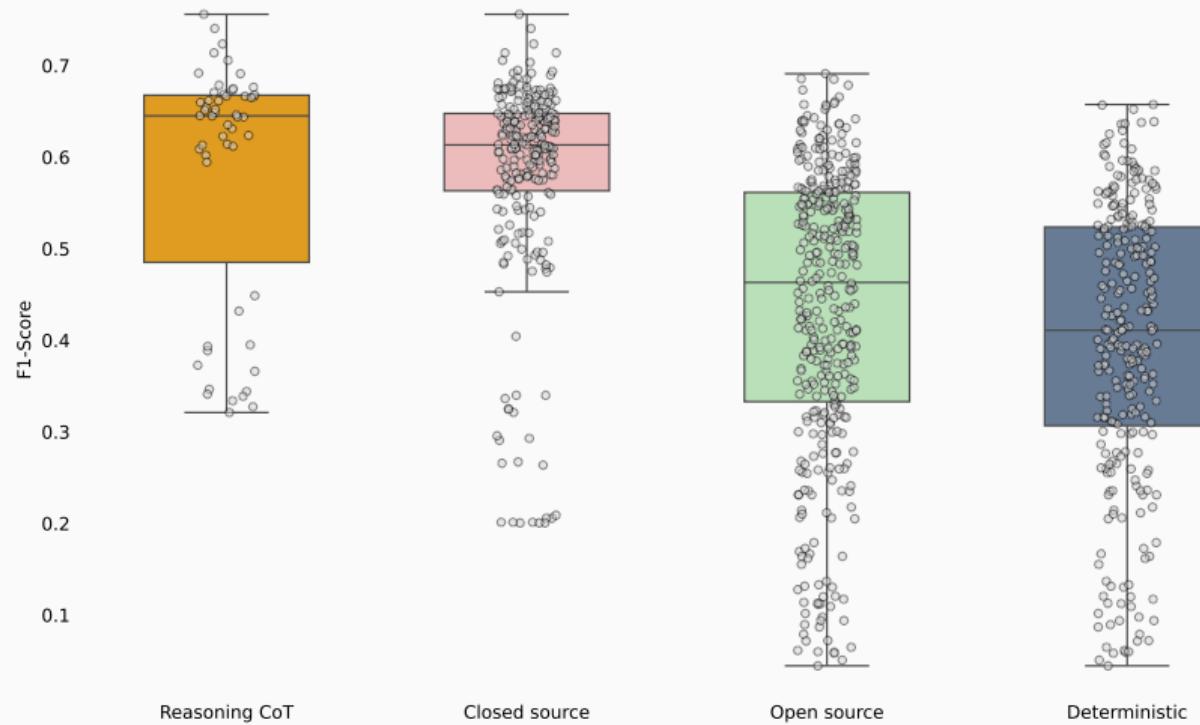
**Mid-Range
Performance**

Gemini 1.5 series,
GPT4o-mini, Mis-
tral models, Qwen 2.5

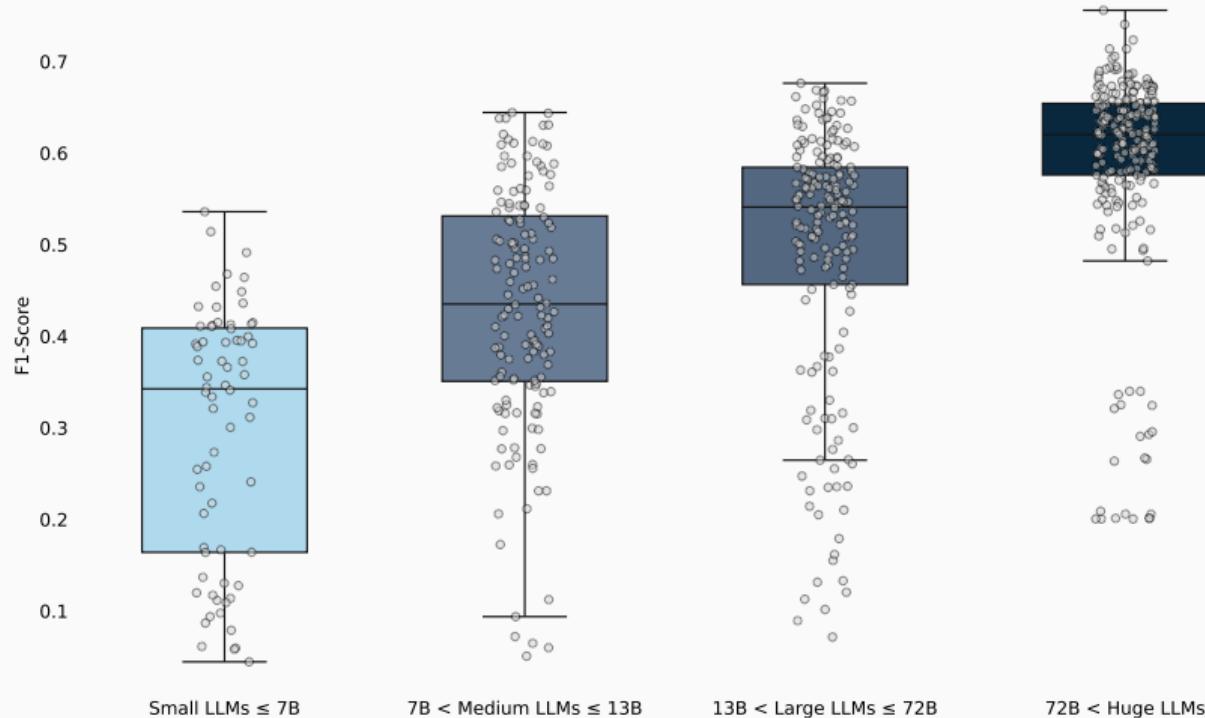
Accuracy Across Reasoning, Source and Deterministic Setup



F1-Score Across Reasoning, Source and Deterministic Setup



F1-Score Across Number of Parameters



Determinants of Performance

	Model I	Model II	Model III	Model IV	Model V
Reasoning CoT	0.462*** (0.031)	0.323*** (0.096)	0.296*** (0.092)	0.359*** (0.086)	0.357*** (0.086)
Open source LLMs		-0.637*** (0.056)	-0.205*** (0.076)	-0.128* (0.072)	-0.126* (0.071)
Deterministic setup			-0.591*** (0.075)	-0.328*** (0.076)	-0.330*** (0.075)
Constant	-0.112*** (0.105)	0.287*** (0.045)	0.291*** (0.043)	-0.476*** (0.092)	-0.285** (0.112)
Parameters	No	No	No	Yes	Yes
Language FE	No	No	No	No	Yes
<i>N</i>	620	620	620	620	620
τ	0.734	0.665	0.633	0.592	0.590
I^2	99.04%	98.84%	98.72%	98.54%	98.53%
R^2	0.029	0.203	0.277	0.368	0.372

Next Steps

1. Fine-Tuning

Comparing with Fine-Tuned Models

Language	Best LLM	F1-Score	Fine-Tuned	F1-Score	Δ Val	Δ Best LLM
Danish	GPT-4.5	0.679	Babel Machine	0.925	+0.065	+0.246
Dutch	o1	0.724	Babel Machine	0.906	+0.066	+0.182
English	o1	0.706	Babel Machine	0.869	-0.031	+0.163
French	o1	0.714	Babel Machine	0.821	-0.029	+0.107
Hungarian	GPT.4-5	0.672	Babel Machine	0.751	-0.099	+0.079
Italian	o1	0.675	Babel Machine	0.930	+0.120	+0.255
Portuguese	o1	0.651	Babel Machine	0.867	-0.063	+0.216
Spanish	o4-mini	0.756	Babel Machine	0.916	+0.066	+0.160

Note. All estimates are weighted F1-scores obtained on our fixed held-out test set. The columns Δ Val and Δ Best LLM indicate: (i) the change relative to the best result on the model's own validation set; and (ii) the change relative to the strongest zero-shot LLM, respectively.

How much data leakage is present here? Probably something, so that the results may be inflated. However, the potential advantages of fine-tuning still seem to outweigh in-context learning, even for BERT-like models. * Babel Machine by [Sebők et al. \(2024\)](#).

Fine-Tuned BERTs



Fine-tuned XLM-RoBERTa

F1 validation 0.819

F1 held-out set 0.810

<https://doi.org/10.57967/hf/6863>



Fine-tuned ModernBERT

F1 validation 0.831

F1 held-out set 0.809

<https://doi.org/10.57967/hf/6864>

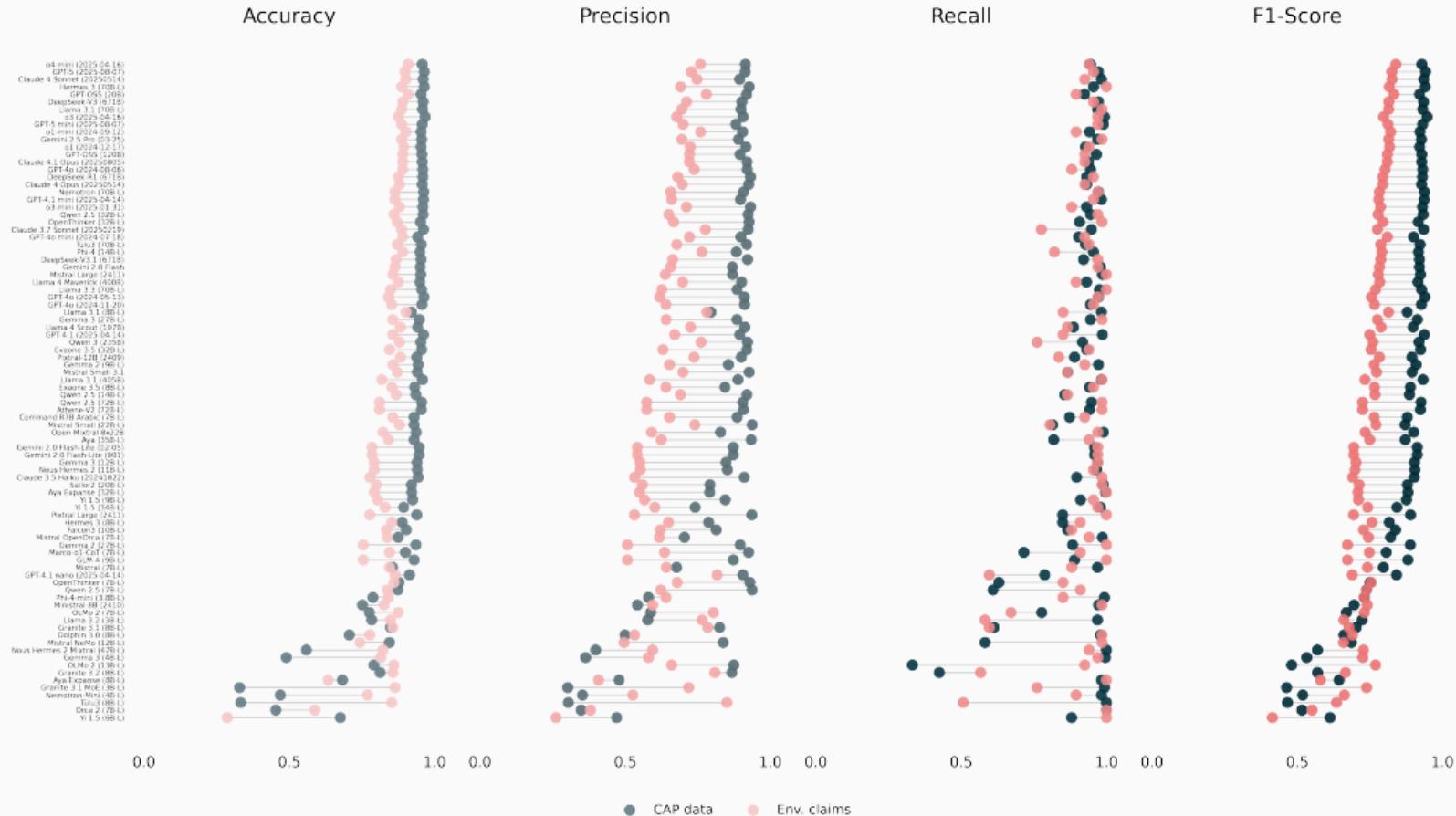
2. Simplify the Task



B González-Bustamante and N van der Zwan

Processing 400 annual reports from leading asset owners, including pension funds, insurance companies and sovereign wealth funds.

Benchmarking nearly 100 contemporary LLMs to select those that best balance accuracy, openness and cost-effectiveness. We used environmental and energy CAP data and environmental claims from [ClimateBERT](#).



Takeaways

Takeaways

+8.9%

**Reasoning
Advantage**

Chain-of-thought capabilities boost F1-score by almost 9 points

-7.8%

**Reproducibility
Penalty**

Deterministic deployment reduces F1-score by about 8 points, but ensures consistency

$\geq 10\%$

**Fine-Tuning
Advantage**

Supervised transformers outperform zero-shot LLMs by 10+ points

We did not find  openness penalty

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with  and some 
 Download the latest version from [GitHub](#) or [Zenodo](#)